Harvard University

Harvard University Biostatistics Working Paper Series

Year 2010 *Paper* 112

Graphical Procedures for Evaluating Overall and Subject-Specific Incremental Values from New Predictors with Censored Event Time Data

Hajime Uno* Tianxi Cai†

Lu Tian‡ L. J. Wei**

*Dana Farber Cancer Institute, huno@hsph.harvard.edu

†Harvard University, tcai@hsph.harvard.edu

‡Stanford University School of Medicine, lutian@stanford.edu

**Harvard University, wei@hsph.harvard.edu

GRAPHICAL PROCEDURES FOR EVALUATING OVERALL AND SUBJECT-SPECIFIC INCREMENTAL VALUES FROM NEW PREDICTORS WITH CENSORED EVENT TIME DATA

Hajime Uno[1,2], T. Cai[2], L. Tian[3], and L.J. Wei[2]

[1]*Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute, Boston, Massachusetts 02115, U.S.A.*

[2]*Department of Biostatistics, Harvard University, Boston, MA 02115, U.S.A*

[3]*Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California 94305, U.S.A*

Summary

Quantitative procedures for evaluating added values from new markers over a conventional risk scoring system for predicting event rates at specific time points have been extensively studied. However, a single summary statistic, for example, the area under the receiver operating characteristic curve or its derivatives, may not provide a clear picture about the relationship between the conventional and the new risk scoring systems. When there are no censored event time observations in the data, two simple scatterplots with individual conventional and new scores for "cases" and "controls" provide valuable information regarding the overall and the subject-specific level incremental values from the new markers. Unfortunately, in the presence of censoring, it is not clear how to construct such plots. In this paper, we propose a nonparametric estimation procedure for the distributions of the differences between two risk scores conditional on the conventional score. The resulting quantile curves of these differences over the subject-specific conventional score provide extra information about the overall added value from the new marker. They also help us to identify a subgroup of future subjects who need the new predictors, especially when there is no unified utility function available for cost-risk-benefit decision making. The procedure is illustrated with two data sets. The first is from a well-known Mayo Clinic PBC liver study. The second is from a recent breast cancer study on evaluating the added value from a gene score, which is relatively expensive to measure compared with the routinely used clinical biomarkers for predicting the patient's survival after surgery.

Keywords: *Discriminant analysis; Nonparametric function estimation; Prediction; Receiver operating characteristic curve.*

1. INTRODUCTION

For a binary phenotypic outcome, numerical and graphical methods for evaluating an overall incremental value from a new set of markers over a conventional risk scoring system have been extensively studied (Bamber, 1975; Zhou et al., 2002; Pepe, 2003; Pepe et al., 2004; Greenland & O'Malley, 2005; Ware, 2006; Pencina et al., 2008). Novel generalizations of these procedures to handle censored event time data have also been proposed (Hanley & McNeil, 1982; Harrell et al., 1996; D'Agostino et al., 1997; Pencina & D'Agostino, 2004; Heagerty and Zheng, 2005; Cook et al., 2006; Cai and Cheng, 2008; Uno et al, 2009). Evaluating the added value from the new markers with an overall summary measure is an important first step for establishing a prediction rule. On the other hand, even when the new markers have either an impressive or no meaningful overall incremental value, the next critical step is to identify patients who would or would not need the additional markers for better prediction via their conventional risk scores. Unfortunately, relatively little effort has been made for establishing a systematic, analytic procedure for such "subgroup analysis" in the statistical or medical literature (D'Agostino, 2006). Recently, Tian et al. (2009) proposed a procedure for this type of subject-specific level analysis by controlling a pre-specified simultaneous inference error rate. However, their proposal does not incorporate censoring and depends heavily on the choice of the utility function, a weighted average between the false positive and negative rates.

For binary outcomes, simple scatterplots of individual conventional risk scores vs. new ones provide valuable information about an overall and also personalized-level incremental values of the new markers (Gu & Pepe, 2009). For example, in selecting patients with advanced or end-stage primary biliary cirrhosis, PBC, for orthotopic liver transplantation, five patients' baseline covariates, namely age, albumin, bilirubin, edema and prothrombin time, were identified to be important predictors for the patient's survival based the data from a Mayo Clinic study (Dickson et al., 1989; Fleming and Harrington, 1991, pp. 160). Suppose that we would like to know the added value from the bilirubin measure over the other four variables with respect to prediction

of 5-year survival based on observations from 416 patients with complete information on those predictors. To this end, we first obtain a risk score based on these four variables without bilirubin,

$$0.29 \times (\text{age}/10) - 3.49 \times \log(\text{albumin}) + 1.33 \times \text{edema} + 3.07 \times \log(\text{prothrombin time}), \quad (1.1)$$

by fitting the data with a simple additive Cox model using partial likelihood estimation procedure (Cox, 1972). Based on (1.1) and the standard Breslow estimator for the baseline cumulative hazard, we obtain individual patients' 5-year cumulative mortality risk, denoted by $p_{1i}, i = 1, \cdots, 416$. Next, we fit the data using another additive Cox model with all five covariates including bilirubin. The resulting risk score is

$$0.40 \times (\text{age}/10) - 2.51 \times \log(\text{albumin}) + 0.86 \times \log(\text{bilirubin}) +$$

$$0.90 \times \text{edema} + 2.4 \times \log(\text{prothrombin time}). \quad (1.2)$$

Let p_{2i} denote the ith individual five-year mortality rate based on (1.2).

In the PBC dataset, there are 196 censored survival observations by Year 5 and 114 patients died during this time period. Figure 1(a) shows the scatterplot of p_{1i} vs. the difference $(p_{2i} - p_{1i})$ for those 114 observable "cases." The majority of those black dots in the figure are above the horizontal line, indicating that globally the bilirubin provides extra information about the 5-year mortality rate for those "cases." Moreover, for a subject with p_1 between 0.2 and 0.6, the corresponding p_2 tends to be substantially higher. Figure 1(b) shows the scatterplot for the observable "controls," who survived and were still under follow-up by Year 5. Here, most of p_2 tend to be smaller than their p_1, indicating that bilirubin has an overall incremental value. At the personalized level, it appears that for the survived patients whose conventional risk scores are between 0.15 and 0.35, bilirubin provides nontrivial improvement for predicting survival beyond 5 years. If there were no censored observations in the data, the scatterplots in Figure 1, coupled with the standard lowess curves for the scatter diagram (dark curves), would provide a valuable tool for quantifying global and subject-specific level performance using Model (1.2)

with bilirubin. Unfortunately, for the present example, the number of censored observations is substantial and it is not clear how to construct valid plots like Figure 1.

In this paper, with censored survival data we propose a nonparametric procedure to consistently estimate quantiles of the distributions of the difference $(p_2 - p_1)$ given p_1 for cases and controls. The resulting quantile curves are then presented using a similar configuration to Figure 1. The new method is derived under a more general setting. Here, a case is defined as the survival time being in a time interval I_1, while a control is defined as the survival time in an interval I_0, where I_1 is entirely on the left hand side of I_0. By repeating the analysis with various pairs of I_0 and I_1, one may find, for example, that the new predictors are not useful when these two intervals are widely separated (for instance, short- vs. long-term survival), but may have substantial incremental values when these two intervals are relatively close. This type of finding can be quite informative for cost-benefit decision making. The new procedure is illustrated with the above Mayo Clinic data and also with the data set from a breast cancer study to evaluate the additional prediction ability based on a new gene risk score on top of conventional clinical markers. The second example is particularly interesting due to the fact that it is relatively expensive to measure the gene score compared with clinical markers, which are routinely obtained after patients' surgery for breast cancer.

2. ESTIMATING THE DISTRIBUTION OF THE NEW RISK SCORE CONDITIONAL ON THE OLD RISK SCORE

Let T be the time to an event of interest and Z be its corresponding vector of baseline covariates. For the two specific time intervals $I_1 \in [t_1, t_2)$ and $I_0 \in [t_3, t_4)$ discussed in Section 1, suppose that for a given Z we are interested in estimating the risk of a case:

$$\text{pr}(T \in I_1 \mid Z)/\{\text{pr}(T \in I_0 \mid Z) + \text{pr}(T \in I_1 \mid Z)\}. \tag{2.1}$$

Let U and V be two vectors, which are functions of Z. Here, U is a function of conventional markers only, but V is a function of both conventional and new predictors. One of the questions

is how to identify patients with U, who may need V for better prediction of (2.1). This is a particularly important question when it is costly or invasive to measure the new markers. Often, the event time T may be censored by a censoring variable C. Assume that C is independent of T and Z. Let $G(\cdot)$ be the survival function of C. Moreover, let the binary variable $E = 0$, if $T \in I_0$; $= 1$, if $T \in I_1$. Note that one can assign an arbitrary value (other than 0 or 1) for E when T is outside of these two time intervals. Now, let $\{(T_i, C_i, E_i, Z_i, U_i, V_i), i = 1, \ldots, n\}$ be n independent copies of (T, C, E, Z, U, V). For T_i, we observe $\{X_i, \Delta_i\}$, where $X_i = \min(T_i, C_i)$, and $\Delta_i = 1$, if $X_i = T_i$, and 0, otherwise, $i = 1, \cdots, n$. Due to potential censoring, the binary variable E may not be observable.

To construct a risk score system with U, let us consider the standard Cox proportional hazards model with the risk score $\beta' U$, where β is an unknown vector of regression parameters. With the above observed data, let $\hat{\beta}$ be the maximum partial likelihood estimator for β. In practice, this semi-parametric model is simply an approximation of the "true" model. Under a mild condition, $\hat{\beta}$ converges to a constant, as $n \to \infty$ (Hjort,1992), regardless of the adequacy of the Cox model. This property is critical for developing our new procedure. Similarly, for V, we fit the data with another additive Cox's model with the risk score $\gamma' V$. Let $\hat{\gamma}$ be the corresponding estimator for γ.

Now, consider an independent future subject from the same study population whose $(T, E, Z, U, V) = (T^0, E^0, Z^0, U^0, V^0)$. To estimate (2.1) with U^0, let $\hat{p}_1(U^0)$ be the estimator for (2.1) constructed from the Breslow estimator for the underlying cumulative hazard function of the above Cox's model and $\hat{\beta}' U^0$. Explicitly, letting $\hat{\Lambda}_1(\cdot)$ denote the Breslow estimator, then $\hat{p}_1(U^0)$ is

$$\frac{\exp\{-\hat{\Lambda}_1(t_1)e^{\hat{\beta}'U^0}\} - \exp\{-\hat{\Lambda}_1(t_2)e^{\hat{\beta}'U^0}\}}{\exp\{-\hat{\Lambda}_1(t_1)e^{\hat{\beta}'U^0}\} - \exp\{-\hat{\Lambda}_1(t_2)e^{\hat{\beta}'U^0)}\} + \exp\{-\hat{\Lambda}_1(t_3)e^{\hat{\beta}'U^0}\} - \exp\{-\hat{\Lambda}_1(t_4)e^{\hat{\beta}'U^0)}\}}.$$

Similarly, let $\hat{p}_2(V^0)$ be the corresponding estimator via the covariate vector V^0. To compare these two predictors, let $\hat{D}(Z^0) = \hat{p}_2(V^0) - \hat{p}_1(U^0)$. Note that to make *overall* comparisons between models with U and V, one may estimate the distribution of $\hat{D}(Z^0)$ given $E^0 = e$, where

e is either 0 or 1. If V has an overall added value over U, one would expect that for $e = 1$, that is, for those future subjects with $T^0 \in I_1$, $\hat{D}(Z^0)$ has more positive mass, and if $e = 0$, $\hat{D}(Z^0)$ has more negative mass. Recently various analytic methods based on the distributions of $\hat{D}(Z^0)$ for cases and controls were proposed, for example, by Pencina et al. (2008), Gu & Pepe (2009) and Uno et al. (2009) to summarize the overall incremental value of the new markers.

In this paper, we are also interested in the subject-specific level evaluation for the incremental values, that is, estimating the distribution of $\hat{D}(Z^0)$ conditional on $E^0 = e$ and $\hat{p}_1(U^0) = p$, where e is either 0 or 1, and p belongs to $\mathcal{J} = [\jmath_l, \jmath_r]$ is a strictly inner subset of the support of $\hat{p}_1(U^0)$. Let $q_{\tau e}(p)$ be the τth conditional quantile of the above distribution, for $0 < \tau < 1$. To estimate $q_{\tau e}(p)$, we utilize a nonparametric quantile regression estimation technique by letting the quantile of $\hat{p}_2(V^0)$ be *locally* linear in $\hat{p}_1(U^0)$ (Yu & Jones, 1998). Specifically, without censored observations, for any given p, we minimize the following objective function with respect to a and b,

$$\sum_{i=1}^{n} I(E_i = e) K_h \left\{\psi(\hat{p}_1(U_i)) - \psi(p)\right\} \rho_\tau(\psi(\hat{p}_2(V_i)) - a - b[\psi\{\hat{p}_1(U_i)\} - \psi(p)]), \tag{2.2}$$

where $K_h(x) = K(x/h)/h$, $K(\cdot)$ is a symmetric probability density function, h is a bandwidth such that $h = O(n^{-\nu})$ with $\nu \in (1/2, 1/5]$ and $\rho_\tau(x)$ is the check function, which is τx if $x \geqslant 0$, and is $(\tau - 1)x$ if $x < 0$. Here, we choose a proper transformation $\psi\{\hat{p}_1(U)\}$ of $\hat{p}_1(U)$ to improve smoothing, where $\psi(\cdot) : (0, 1) \to (-\infty, \infty)$ is a known, non-decreasing function (Wand et al., 1991; Park et al., 1997). For example, one may let $\psi(p) = \log\{-\log(1 - p)\}$. Let the minimizer of (2.2) be $\hat{a}$ and $\hat{b}$. Then let

$$\hat{q}_{\tau e}(p) = \psi^{-1}(\hat{a}) - p \tag{2.3}$$

be an estimator for $q_{\tau e}(p)$.

Since E may not be observable in (2.2), we replace $I(E = e)$ by $I(E^\dagger = e)$, with an inverse probability weighting technique, where $E^\dagger$ is 1 if $X \in I_1$; 0 if $X \in I_0$. Specifically, let $\hat{G}(\cdot)$ denote the Kaplan Meier estimator of $G(\cdot)$ and let η be a pre-specified time point such that $G(\eta) > 0$. The choice of the weight depends of the choice of I_1 and I_0. For the case where I_0 is an interval

such that $t_4 < \eta$, the weight $\widehat{w}_i = \Delta_i/\hat{G}(X_i)$ for both $I(E_i^\dagger = 1)$ and $I(E_i^\dagger = 0)$. This may be justified heuristically using the argument that $E\{\widehat{w}_i I(E_i^\dagger = e) \mid Z_i, T_i\} \approx E\{I(E = e) \mid Z\}$. For the case when $t_3 < \eta$, and $t_4 = \infty$ for the interval I_0, the weights $\widehat{w}_i = \Delta_i/\hat{G}(X_i)$ and $\widehat{w}_i = 1/\hat{G}(t_3)$ are used for $I(E_i^\dagger = 1)$ and $I(E_i^\dagger = 0)$, respectively. Heuristically this can be justified with the argument that $E\{\widehat{w}_i I(E_i^\dagger = 0) \mid Z_i, T_i\} \approx E\{I(E = 0) \mid Z\}$. This inverse probability weighting adjustment, coupled with (2.2), results in the following minimand:

$$\sum_{i=1}^{n} \widehat{w}_i I(E_i^\dagger = e) K_h \left\{\psi\{\hat{p}_1(U_i)\} - \psi(p)\right\} \rho_\tau(\psi\{\hat{p}_2(V_i)\} - a - b[\psi\{\hat{p}_1(U_i)\} - \psi(p)]). \tag{2.4}$$

Then, the corresponding estimator $\hat{q}_{\tau e}(p)$ for $q_{\tau e}(p)$ is given by (2.3), but with $\hat{a}$ being a minimizer of (2.4) with respect to a and b. In the Appendix, we show that for each fixed τ, $\sup_{p\in\mathcal{J}} |\hat{q}_{\tau e}(p) - q_{\tau e}(p)| \to 0$, in probability as $n \to \infty$.

In practice, it is important to know how to choose the smooth parameter h in the above nonparametric estimation. To this end, we consider a commonly used K-fold cross-validation procedure. Specifically, we randomly partition the data into K disjoint parts, $\mathcal{I}_1, \cdots, \mathcal{I}_K$. For each k, we use the data not in $\mathcal{I}_k$ to obtain the regression parameter estimators in the above two Cox's models, denoted by $\hat{\beta}_{(-k)}$ and $\hat{\gamma}_{(-k)}$. Moreover, let $\{\hat{p}_{1(-k)}(\cdot), \hat{p}_{2(-k)}(\cdot), \hat{q}_{\tau e(-k)}(\cdot)\}$ denote the respective estimators corresponding to $\{\hat{p}_1(\cdot), \hat{p}_2(\cdot), \hat{q}_{\tau e}(\cdot)\}$ based on data not in $\mathcal{I}_k$. We propose to choose h by minimizing

$$\sum_{k=1}^{K} \sum_{i\in\mathcal{I}_k} \widehat{w}_i I(E_i^\dagger = e) I\{\hat{p}_{1(-k)}(U_i) \in \mathcal{J}\} \rho_\tau \left(\psi\left\{\hat{p}_{2(-k)}(V_i)\right\} - \psi\left[\hat{q}_{\tau e(-k)}\{\hat{p}_{1(-k)}(U_i)\} + \hat{p}_{1(-k)}(U_i)\right]\right). \tag{2.5}$$

In practice, the lower and upper bounds of $\mathcal{J}$ may be chosen as, for example, the 3rd and 97th percentiles of the empirical distribution of $\hat{p}_1(U^0)$.

3. EXAMPLES

First, let us revisit the PBC example discussed in the Introduction Section. Figure 2 gives the Kaplan-Meier estimate with the survival times from 416 patients. Assume that we are

interested in two time intervals, $I_0 = (5, \infty)$(years) and $I_1 = [0, 5]$ (years). On average the 5-year cumulative mortality rate is about 0.3. Here, $\hat{p}_1(U_i)$ is obtained without using bilirubin and $\hat{p}_2(V_i)$ is with bilirubin, $i = 1, \cdots, n$, via the risk scores (1.1) and (1.2) and two working Cox's models. To estimate $q_{\tau e}(\cdot)$, we let ψ in (2.4) be the $\log(-\log)$ function and the kernel function be the standard normal density function. To choose the smooth parameter h, we used 10-fold cross validation scheme with (2.5). For instance, to estimate the median $q_{0.5e}(\cdot)$ for patients who would die by Year 5, the resulting "optimal" h with respect to ψ-scale is 1.6. The interval $\mathcal{J}$ over which we construct the median curves is $(0.10, 0.995)$. Figure 3(a) gives the estimated median curve of $\hat{D}$ over the risk score $\hat{p}_1$ (solid curve). The lower and upper boundaries of the shaded area are the corresponding $25th$ and $75th$ percentile curves. Figure 3(b) gives the plots which are the counterparts for subjects who would survive more than 5 years. In Figure 3(c), we provide the density function estimate of $\hat{p}_1$ score. The majority of patients whose risk scores without using bilirubin are between 0.1 and 0.6. Based on Figure 3, the distributions of $\hat{D}$ for "cases" over the interval $(0.1, 0.995)$ have more positive mass, especially for $\hat{p}_1$ between 0.2 and 0.6. The bilirubin helps greatly for the "controls" when $\hat{p}_1$ is between 0.2 and 0.6, that is, the false positive rate can be drastically reduced with bilirubin. We have also examined extensively the added values of bilirubin for various sets of time intervals I_0 and I_1. In Figure 4, we present the plots of estimated median curves for cases and controls with respect to four different sets of time intervals I_0 and I_1. If bilirubin were not routinely measured for evaluating liver function clinically, one would recommend its usage for future subjects whose "conventional" scores were between 0.2 and 0.6. Note that we cannot estimate the medians well for controls beyond 0.6 with this set of data.

Next, we use a more interesting example to illustrate a scenario in which a non-trivial cost is associated with measuring a new marker. The data set used for our illustration is from a breast cancer study to evaluate a new genetic marker, "wound-response gene expression signature," for predicting patients' survival (Chang et al., 2005). For each study patient, this gene score was

derived from the microarray gene expression data. Here, the data set consists of 295 breast cancer patient files. Each file is composed of a patient's clinical outcomes (metastasis/death or censoring time), the gene score, and conventional baseline variables collected at time of surgery, including age, tumor diameter, number of positive lymph-nodes, tumor grade, vascular invasion, estrogen receptor status, chemo/hormonal therapy or not, and mastectomy or breast conserving surgery. The data are available at http://microarray-pubs.stanford.edu/wound_NKI/explore.html, which were collected at the Netherlands Cancer Institute by van't Veer et al. (2002) and van de Vijver et al. (2002). The median follow-up time for those 295 patients is 6.7 years and the range is 0.05 to 18.3 years. The gene score (the so-called Dutch 70) created by the aforementioned Dutch scientists is different from that proposed by Chang et al. (2005). Here, we are interested in quantifying the added value from the gene score by Chang et al. over the above conventional clinical predictors. To this end, we fit the data with two working Cox models, one with gene score and the other without. The regression coefficient estimates are given in Table 1. Note that the gene score is statistically significant. The Kaplan-Meier curve of the survival times from these 295 patients is given in Figure 5. First, assume that we are interested in $I_0 = (10, \infty)$ (years) and $I_1 = (0, 10]$ (years). For our analysis, we used the standard normal kernel and the $\log(-\log)$ as the ψ function in the nonparametric estimation of the quantiles. Moreover, we used 10-fold cross validation procedure to choose h. For example, for estimating the medians, the optimal h for "cases" is 3.25 with respect to the ψ-scale and $\mathcal{J}$ is $(0.15, 0.85)$. Figure 6(a) gives the median curve (solid curve) and the bands whose boundaries are the 0.25 and 0.75 quantile curves for "cases." The x-axis is the score without using gene expression data. Figure 6(b) gives the counterparts for "controls," those subjects who would survive beyond 10 years. The density function estimate of $\hat{p}_1$ is given in Figure 6(c). The "conventional scores" of the majority of patients in this population are between 0.2 and 0.75. Note that the median curve is in the positive (negative) side for cases (controls). The improvement from the gene score, however, is quite modest uniformly over the conventional score. Since it is relatively expensive to measure the gene score compared with

the routinely obtained clinical marker values, it is not clear from cost-benefit view if we should measure the gene score for any future patient. In Figure 7, we present plots of estimated median curves for cases and controls with respect to various sets of time intervals I_0 and I_1. Again, there seems no obvious gain from measuring gene score for predicting survival.

4. REMARKS

If a study is designed for evaluating the incremental value from new predictors with respect to a specific set of time intervals I_0 and I_1, a global Cox model may not be appropriate for establishing the risk scores due to the fact that the resulting regression coefficient estimates reflect an average covariate effect over the entire study time. For this case, we may use, for example, the logistic regression for modeling the probability of a binary variable with two events $\{T \in I_1\}$ and $\{T \in I_0\}$ with predictors and use the technique developed by Uno et al. (2007) to obtain the risk scores. Then with the same argument in the present paper, nonparametric function estimates for conditional quantiles can be obtained accordingly. When there is no pre-specified set of time intervals of interest, one may use the Cox models to obtain unified scores $\hat{\beta}'U$ and $\hat{\gamma}'V$ first. However, it is important to note that these two scoring systems may not be comparable since we fit the data with two different models. Therefore, in this paper we convert the Cox scores to their risk counterparts with respect to a given paired I_0 and I_1 to evaluate the incremental values. By considering various sets of I_0 and I_1 in our analysis, one may identify when the new markers have practically meaningful added values for prediction. On the other hand, it is not clear how to utilize the Cox scores directly to perform such subject-level analysis without discretizing the continuous study follow-up time.

If the conventional scoring system is well-established, one may not need to fit the current data with the conventional markers. However, for this situation we recommend examining closely whether the present study population is similar to that from which the conventional score was constructed.

The graphical method presented here can also be utilized as a *quantitative* way to assess

relative merits of two proposed models for fitting survival data. Unlike the lack of fit tests for model checking or a single summary statistic such as the likelihood ratio, the plots in Figures 4 and 7 with different sets of I_0 and I_1 provide much more information to help us to choose an appropriate model with respect to cost-benefit decision making.

It is important to note that the parametric or semi-parametric models used for constructing the risk scores are simply approximations for the true models. If the "old" model does not fit the data well, it is difficult if not impossible to determine whether the improvement from the "new" model is the incremental value from the new predictors or a better model fitting.

5. Appendix

Throughout, unless noted otherwise, we use the notation $\simeq$ to denote equivalence up to $o_p(1)$ uniformly in p, $\lesssim$ to denote being bounded above up to a universal constant, and $\dot{\mathcal{F}}(x)$ to denote $d\mathcal{F}(x)/dx$ for any function $\mathcal{F}$.

We use $\mathbb{P}_n$ and $\mathbb{P}$ to denote expectation with respect to the empirical probability measure of $\{(X_i, \delta_i, Z_i), i = 1, \cdots, n\}$ and the probability measure of (X, δ, Z) respectively. Similarly $\mathbb{G}_n = n^{\frac{1}{2}}(\mathbb{P}_n - \mathbb{P})$. Let $\widehat{\theta} = (\widehat{\theta}_1', \widehat{\theta}_2')'$, $\widehat{\theta}_1 = (\log \hat{\Lambda}_1(t_1), \log \hat{\Lambda}_1(t_2), \log \hat{\Lambda}_1(t_3), \log \hat{\Lambda}_1(t_4), \widehat{\beta}')'$, $\widehat{\theta}_2 = (\log \hat{\Lambda}_2(t_1), \log \hat{\Lambda}_2(t_2), \log \hat{\Lambda}_2(t_3), \log \hat{\Lambda}_2(t_4), \widehat{\gamma}')'$, where $\hat{\Lambda}_1(\cdot)$ and $\hat{\Lambda}_2(\cdot)$ are the estimated cumulative hazard functions based on the models with U and with V respectively. Note that $\widehat{p}_1(U^0) = g(\widehat{\theta}_1'\vec{U}_1^0, \widehat{\theta}_1'\vec{U}_2^0, \widehat{\theta}_1'\vec{U}_3^0, \widehat{\theta}_1'\vec{U}_4^0)$ and $\widehat{p}_2(V^0) = g(\widehat{\theta}_2'\vec{V}_1^0, \widehat{\theta}_2'\vec{V}_2^0, \widehat{\theta}_2'\vec{V}_3^0, \widehat{\theta}_2'\vec{V}_4^0)$, where $g(x_1, x_2, x_3, x_4) = \{\exp(-e^{x_1}) - \exp(-e^{x_2})\}/\{\exp(-e^{x_1}) - \exp(-e^{x_2}) + \exp(-e^{x_3}) - \exp(-e^{x_4})\}$ and for any vector x, $\vec{x}_1 = (1, 0, 0, 0, x')'$, $\vec{x}_2 = (0, 1, 0, 0, x')'$, $\vec{x}_3 = (0, 0, 1, 0, x')'$, and $\vec{x}_4 = (0, 0, 0, 1, x')'$.

To establish the consistency of the proposed estimator, we assume that $h = O(n^{-\nu})$ with $1/2 > \nu > 1/5$ and $\widehat{\theta}$ converges in probability to a deterministic vector $\theta_0 = (\theta_{10}', \theta_{20}')'$. Let $\bar{p}_1(U^0) = g(\theta_{10}'\vec{U}_1^0, \theta_{10}'\vec{U}_2^0, \theta_{10}'\vec{U}_3^0, \theta_{10}'\vec{U}_4^0)$, $\bar{p}_2(V^0) = g(\theta_{20}'\vec{V}_1^0, \theta_{20}'\vec{V}_2^0, \theta_{20}'\vec{V}_3^0, \theta_{20}'\vec{V}_4^0)$, $\bar{p}_{\mathfrak{o}i} = \bar{p}_1(U_i)$, $\bar{p}_{\mathfrak{n}i} = \bar{p}_2(V_i)$, $\mathcal{P}_0(p) = P(E_i = 0 \mid \bar{p}_{\mathfrak{o}i} = p)$, $\mathcal{P}_1(p) = P(E_i = 1 \mid \bar{p}_{\mathfrak{o}i} = p)$ and $\zeta_{ep}(x)$ denote the conditional density of $\psi(\bar{p}_{\mathfrak{n}i})$ given $\bar{p}_{\mathfrak{o}i} = p$ and $E_i = e$ which is assumed to be continuously differentiable. We assume that $\xi(x)$, the density function of $\psi(\bar{p}_{\mathfrak{o}i})$, is continuously differentiable

with bounded derivatives and bounded away from zero for $x \in \mathcal{J}$. We also assume that U and V are bounded, θ_0 is an interior point of a compact set Ω. Furthermore, $\widehat{\theta}$ is a regular estimator of θ_0 with

$$n^{1/2}(\widehat{\theta} - \theta_0) = n^{1/2} \begin{bmatrix} \widehat{\theta}_1 - \theta_{10} \\ \widehat{\theta}_2 - \theta_{20} \end{bmatrix} = n^{-1/2} \sum_{i=1}^{n} \begin{bmatrix} \mathcal{W}_{1i} \\ \mathcal{W}_{2i} \end{bmatrix} + o_p(1). \tag{5·1}$$

where $\mathcal{W}_{1i} = \mathcal{W}_1(X_i, \delta_i, U_i)$ and $\mathcal{W}_{2i} = \mathcal{W}_2(X_i, \delta_i, V_i)$ for some deterministic function $\mathcal{W}_1$ and $\mathcal{W}_2$. See Uno et al (2007) and Cai et al (2009) for details on establishing the above asymptotic properties. Furthermore, we note that $\sup_{t \leqslant t_0} |n^{1/2}\{\widehat{G}(t) - G(t)\}| = O_p(1)$ (Kalbfleish and Prentice, 2002). It follows that

$$\|\widehat{\theta} - \theta_0\| + \sup_{t \leqslant t_0} |\widehat{G}(t) - G(t)| = O_p(n^{-1/2}). \tag{5·2}$$

It follows that $\sup_{d, p \in [\jmath_l, \jmath_r]} |P\{\hat{D}(Z^0) \leqslant d \mid E^0 = e, \widehat{p}_1(U^0) = p\} - P\{\bar{D}(Z^0) \leqslant d \mid E^0 = e, \bar{p}_1(U^0) = p\}| = O_p(n^{-1})$ and thus $|q_{\tau e}(p) - \bar{q}_{\tau e}(p)| = O_p(n^{-1})$, where $\bar{q}_{\tau e}(p)$ is the τth quantile of $\bar{D}(Z^0)$ given $E^0 = e$ and $\bar{p}_1(U^0) = p$. Thus to establish the consistency of $\hat{q}_{\tau e}(p)$ for $p \in [\jmath_l, \jmath_r]$, it suffices to show that $\hat{q}_{\tau e}(p)$ is uniformly consistent for $\bar{q}_{\tau e}(p)$.

For the ease of notation, we next establish the consistency of the conditional median $\hat{q}_{0.5e}(p)$ for the case with $e = 1$ and note that similar arguments can be used for other quantiles. For any given p, let $a(p) = \bar{q}_{\tau e}(p) + p$ and $b(p) = \dot{a}(p)$ for $\tau = 0.5$ and $e = 1$, $\widehat{\mathcal{E}}_i(p) = \psi(\widehat{p}_{\mathfrak{o}i}) - \psi(p)$, $\{\widehat{a}(p), \widehat{b}(p)\}$ be the minimizer of

$$\widehat{L}(a, b; p) = n^{-1} \sum_{i=1}^{n} \widehat{w}_i E_i^{\dagger} K_h\{\widehat{\mathcal{E}}_i(p)\} \left| \psi(\widehat{p}_{\mathfrak{n}i}) - a - b\widehat{\mathcal{E}}_i(p) \right|$$

$$\text{and} \qquad \widehat{\boldsymbol{\varepsilon}}(p) = \begin{bmatrix} \widehat{\varepsilon}_a(p) \\ \widehat{\varepsilon}_b(p) \end{bmatrix} = \begin{bmatrix} \widehat{a}(p) - a(p) \\ h\{\widehat{b}(p) - b(p)\} \end{bmatrix}.$$

Our objective is to show that $\sup_p |\widehat{\boldsymbol{\varepsilon}}(p)| \to 0$ in probability as $n \to \infty$. To this end, we note that for any given p, $\widehat{\boldsymbol{\varepsilon}}(p)$ is the minimizer of the objective function $\widehat{\mathbb{L}}(\boldsymbol{\varepsilon}; p) = 0$, where $\boldsymbol{\varepsilon} = (\varepsilon_a, \varepsilon_b)'$,

$$\widehat{\mathbb{L}}(\boldsymbol{\varepsilon}; p) = n^{-1} \sum_{i=1}^{n} \widehat{w}_i E_i^{\dagger} K_h\{\widehat{\mathcal{E}}_i(p)\} \left| \psi(\widehat{p}_{\mathfrak{n}i}) - \mathcal{G}\{\boldsymbol{\varepsilon}, p; \psi(\widehat{p}_{\mathfrak{o}i}), h\} \right|,$$

$$\text{and} \quad \mathcal{G}(\boldsymbol{\varepsilon}, p; y) = a(p) + b(p)\{y - \psi(p)\} + \varepsilon_a + \varepsilon_b h^{-1}\{y - \psi(p)\}.$$

It suffices to show that $\widehat{\mathbb{L}}(\boldsymbol{\varepsilon};p)$ is uniformly consistent for

$$\mathbb{L}(\boldsymbol{\varepsilon};p)=\xi\{\psi(p)\}\mathcal{P}_1(p)\int\int|u-\{a(p)+\varepsilon_a+\varepsilon_b v\}|\,K(v)\zeta_{1p}(u)dvdu$$

To this end, we note that $|\widehat{\mathbb{L}}(\boldsymbol{\varepsilon};p)-\mathbb{L}(\boldsymbol{\varepsilon};p)|\leqslant\mathscr{E}_1(\boldsymbol{\varepsilon};p)+\mathscr{E}_2(\boldsymbol{\varepsilon};p)+\mathscr{E}_3(\boldsymbol{\varepsilon};p)$, where

$$\begin{aligned}
\mathscr{E}_1(\boldsymbol{\varepsilon};p)&=n^{-1}\sum_{i=1}^{n}|\widehat{w}_i-w_i|\,E_i^{\dagger}K_h\{\widehat{\mathcal{E}}_i(p)\}\,|\psi(\widehat{p}_{\mathfrak{n}i})-\mathcal{G}\{\boldsymbol{\varepsilon},p;\psi(\widehat{p}_{\mathfrak{o}i}),h\}|\\
\mathscr{E}_2(\boldsymbol{\varepsilon};p)&=\big|\mathbb{P}_nK_h\{\mathcal{E}(p)\}wE^{\dagger}\,|\psi(\bar{p}_{\mathfrak{n}})-\mathcal{G}\{\boldsymbol{\varepsilon},p;\psi(\bar{p}_{\mathfrak{o}}),h\}|-\mathbb{L}(\boldsymbol{\varepsilon};p)\big|\\
\mathscr{E}_3(\boldsymbol{\varepsilon};p)&=\Big|\mathbb{P}_nwE^{\dagger}\Big[K_h\{\widehat{\mathcal{E}}(p)\}\,|\psi(\widehat{p}_{\mathfrak{n}})-\mathcal{G}\{\boldsymbol{\varepsilon},p;\psi(\widehat{p}_{\mathfrak{o}})\}|\\
&\qquad\qquad-K_h\{\mathcal{E}(p)\}\,|\psi(\bar{p}_{\mathfrak{n}})-\mathcal{G}\{\boldsymbol{\varepsilon},p;\psi(\bar{p}_{\mathfrak{o}})\}|\,\Big]\Big|
\end{aligned}$$

and $w_i=\Delta_i/G(X_i)$, First, following directly from (5·2), $\sup_{\boldsymbol{\varepsilon};p}\mathscr{E}_1(\boldsymbol{\varepsilon};p)=o_p(1)$. Secondly, with the standard arguments used in Bickel & Rosenblatt (1973), $\sup_{\boldsymbol{\varepsilon};p}|\mathscr{E}_2(\boldsymbol{\varepsilon};p)|=O_p\{(nh)^{-\frac{1}{2}}\log(n)\}=o_p(1)$. Lastly, for $\mathscr{E}_3(\boldsymbol{\varepsilon};p)$, we note that from the inequality that $\big|a_1|b_1|-a_2|b_2|\big|\leqslant a_1|b_1-b_2|+|b_2||a_1-a_2|$ for $a_1,a_2>0$,

$$\begin{aligned}
\mathscr{E}_3(\varepsilon,p)&\leqslant\mathbb{P}_nwE^{\dagger}K_h\{\widehat{\mathcal{E}}(p)\}\Big|\psi(\widehat{p}_{\mathfrak{n}})-\mathcal{G}\{\boldsymbol{\varepsilon},p;\psi(\widehat{p}_{\mathfrak{o}})\}-\psi(\bar{p}_{\mathfrak{n}})+\mathcal{G}\{\boldsymbol{\varepsilon},p;\psi(\bar{p}_{\mathfrak{o}})\}\Big|\\
&\quad+\Big|\mathbb{P}_n\Big[K_h\{\widehat{\mathcal{E}}(p)\}-K_h\{\mathcal{E}(p)\}\Big]wE^{\dagger}\,|\psi(\bar{p}_{\mathfrak{n}})-\mathcal{G}\{\boldsymbol{\varepsilon},p;\psi(\bar{p}_{\mathfrak{o}})\}|\,\Big|\\
&\lesssim O_p(h^{-1}n^{-1/2})+\int\int\int\int K_h[\psi\{g(u_1,u_2,u_3,u_4)\}-\psi(p)]\mathbb{H}_n(du_1,du_2,du_3,du_4)
\end{aligned}$$

where $\mathbb{H}_n(u_1,u_2,u_3,u_4)=\mathbb{P}_nwE^{\dagger}\,|\psi(\bar{p}_{\mathfrak{n}})-\mathcal{G}(\boldsymbol{\varepsilon},p;\psi(\bar{p}_{\mathfrak{o}}))|\,\{I(\widehat{\theta}_1'\vec{U}_1\leqslant u_1,\widehat{\theta}_1'\vec{U}_2\leqslant u_2,\widehat{\theta}_1'\vec{U}_3\leqslant u_3,\widehat{\theta}_1'\vec{U}_4\leqslant u_4)-I(\theta_{10}'\vec{U}_1\leqslant u_1,\theta_{10}'\vec{U}_2\leqslant u_2,\theta_{10}'\vec{U}_3\leqslant u_3,\theta_{10}'\vec{U}_4\leqslant u_4)\}$. Furthermore, it follows from integration by parts,

$$\begin{aligned}
\sup_{\boldsymbol{\varepsilon};p}\mathscr{E}_3(\boldsymbol{\varepsilon};p)&\lesssim h^{-1}\sup|\mathbb{H}_n(u)|\\
&\lesssim n^{-1/2}h^{-1}\Big|\mathbb{G}_n[wE^{\dagger}\,|\psi(\bar{p}_{\mathfrak{n}})-\mathcal{G}(\boldsymbol{\varepsilon},p;\psi(\bar{p}_{\mathfrak{o}}))|\,\{I(\widehat{\theta}_1'\vec{U}_1\leqslant u_1,\widehat{\theta}_1'\vec{U}_2\leqslant u_2,\widehat{\theta}_1'\vec{U}_3\leqslant u_3,\widehat{\theta}_1'\vec{U}_4\leqslant u_4)-\\
&\quad I(\theta_{10}'\vec{U}_1\leqslant u_1,\theta_{10}'\vec{U}_2\leqslant u_2,\theta_{10}'\vec{U}_3\leqslant u_3,\theta_{10}'\vec{U}_4\leqslant u_4)\}]\Big|+o_p(1).
\end{aligned}$$

Since the class of functions $\{wE^{\dagger}\,|\psi(\bar{p}_{\mathfrak{n}})-\mathcal{G}\{\boldsymbol{\varepsilon},p;\psi(\bar{p}_{\mathfrak{o}})\}|\,I(\theta_1'\vec{U}_1-u_1\leqslant 0,\theta_1'\vec{U}_2-u_2\leqslant 0,\theta_1'\vec{U}_3-u_3\leqslant 0,\theta_1'\vec{U}_4-u_4\leqslant 0):\|\theta_1-\theta_{10}\|\leqslant\delta,u_1,u_2,u_3,u_4,\varepsilon,p\}$, indexed by $(\theta_1,u_1,u_2,u_3,u_4,\varepsilon,p)$, is Donsker, the stochastic process $\mathbb{G}_n[wE^{\dagger}\,|\psi(\bar{p}_{\mathfrak{n}})-\mathcal{G}(\boldsymbol{\varepsilon},p;\psi(\bar{p}_{\mathfrak{o}}))|\,I(\theta_1'\vec{U}_1-u_1\leqslant 0,\theta_1'\vec{U}_2-u_2\leqslant 0,\theta_1'\vec{U}_3-u_3\leqslant 0,\theta_1'\vec{U}_4-u_4\leqslant 0)]$

is stochastically continuous in θ_1. This, coupled with the fact that $\hat{\theta}_1 - \theta_{10} = o_p(1)$, implies that $\sup \left| \mathbb{G}_n[wE^{\dagger} \left| \psi(\bar{p}_{\mathfrak{n}}) - \mathcal{G}(\boldsymbol{\varepsilon}, p; \psi(\bar{p}_{\mathfrak{o}})) \right| \{ I(\widehat{\theta}_1' \vec{U}_1 \leqslant u_1, \widehat{\theta}_1' \vec{U}_2 \leqslant u_2, \widehat{\theta}_1' \vec{U}_3 \leqslant u_3, \widehat{\theta}_1' \vec{U}_4 \leqslant u_4) - I(\theta_{10}' \vec{U}_1 \leqslant u_1, \theta_{10}' \vec{U}_2 \leqslant u_2, \theta_{10}' \vec{U}_3 \leqslant u_3, \theta_{10}' \vec{U}_4 \leqslant u_4) \}] \right| = o_p(1)$. Therefore, $\mathscr{E}_3(\varepsilon, p) = O_p(n^{-1/2} h^{-1}) = o_p(1)$ uniformly in ε and u. It follows that $\sup_{\boldsymbol{\varepsilon}; p} |\widehat{\mathbb{L}}(\boldsymbol{\varepsilon}; p) - \mathbb{L}(\boldsymbol{\varepsilon}; p)| = o_p(1)$. This uniform convergence, coupled with the fact that 0 is the unique minimizer of $\mathbb{L}(\boldsymbol{\varepsilon}, p)$ with respect to $\boldsymbol{\varepsilon}$, suggests that $\sup_p |\widehat{\boldsymbol{\varepsilon}}(p)| = o_p(1)$, which implies the uniform consistency of $\widehat{a}(p)$ for $a(p)$. This, together with $|q_{\tau e}(p) - \bar{q}_{\tau e}(p)| = O_p(n^{-1})$, implies the uniform consistency of $\hat{q}_{\tau e}(p)$ for $q_{\tau e}(p)$ for $\tau = 0.5$ and $e = 1$.

References

Bamber D. (1975), "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *Journal of Mathematical Psychology,* 12, 387–415.

Bickel, P. J. & Rosenblatt, M. (1973), "On some global measures of the deviations of density function estimates (Corr: V3 p1370)," *The Annals of Statistics,* 1, 1017 – 1095.

Cai, T. & Cheng, S. (2008), "Robust combination of multiple diagnostic tests for classifying censored event times," *Biostatistics,* 9, 216–233.

Cai, T., Tian, L., Uno, H. Solomon, S. D. & Wei, L. J. (2010), "Calibrating parametric subject-specific risk estimation," *Biometrika,* to appear.

Chang, H. Y., Nuyten, D. S. A., Sneddon, J. B., Hastie, T., Tibshirani, R., Sørlieb, T., Dai, H., He, Y. D., van't Veer, L. J., Bartelinke, H., van de Rijnj, M., Brownb, P. O. & van de Vijver, M.J. (2005), "Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival," *PNAS,* 102, 3738–43.

Cook, N. R., Buring, J. E., & Ridker, P. M. (2006), "The effect of including C-reactive protein in cardiovascular risk prediction models for women," *Annals of Internal Medicine,* 145, 21 – 29 .

Cox, D. R. (1972), "Regression Models and Life Tables" (with Discussion), *Journal of the Royal Statistical Society,* Ser. B, 34, 187–220.

D'Agostino, R. B. (2006), "Risk prediction and finding new independent prognostic factors," *Journal of Hypertension*, 24, 643–645.

D'Agostino, R. B., Griffith, J. L., Schmidt, C. H., & Terrin, N. (1997), "Measures for evaluating model performance," *Proceedings of the Biometrics Section, Alexandria, VA, U.S.A. American Statistical Association, Biometrics Section: Alexandria, VA.,* 253 – 258

Dickson, E., Fleming, T., Grambsch, P., Fisher, L., and Langworthy, A. (1989), "Prognosis in Primary Biliary Cirrhosis: Model for Decision Making," *Hepatology,* 10, 1–7.

Fleming, T. R. and Harrington, D. P. (1991), *Counting Processes and Survival Analysis,* New York: John Wiley & Sons, Inc.

Gu, W. & Pepe, M. (2009), "Measures to summarize and compare the predictive capacity of markers," *The International Journal of Biostatistics,* 5 (1), 2454 – 2456.

Greenland, P. & O'Malley, P. G. (2005), "When is a new prediction marker useful? A consideration of lipoprotein-associated phospholipase A2 and C-reactive protein for stroke risk," *Archives of Internal Medicine,* 165(21), 2454 – 2456.

Hanley, J. A. & McNeil, B. J. (1982), "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology,* 143, 29 – 36.

Harrell, F. E., Lee, K. L., & Mark, D.B. (1996), "Tutorial in Biostatistics: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in Medicine,* 15, 361–87.

Heagerty, P. J. & Zheng, Y. (2005), "Survival Model Predictive Accuracy and ROC Curves," *Biometrics,* 61, 92–105.

Hjort, N. (1992), "On inference in parametric survival data models," *International Statistical Review,* 60, 355–87.

Kalbfleish, J. D. & Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data* (2nd ed.), New York: John Wiley & Sons, Inc.

Park, B., Kim, W., Ruppert, D., Jones, M., Signorini, D. & Kohn, R. (1997), "Simple transfor-

mation techniques for improved non-parametric regression," *Scandinavian journal ofstatistics,* 24, 145 – 163.

Pencina, M. J. & D'Agostino, R. B. (2004), "Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation," *Statistics in Medicine,* 23, 2109–23.

Pencina, M. J., D'Agostino, R. B. Sr., D'Agostino, R. B. Jr., & Vasan, R. S. (2008), "Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond (with Commentaries & Rejoinder)," *Statistics in Medicine,* 27, 157–212.

Pepe, M. S. (2003), "The statistical evaluation of medical tests for classification and prediction," *Oxford University Press, New York.*

Pepe, M. S., Janes, H., Longton, G., Leisenring, W. & Newcomb P. (2004), "Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker," *American Journal of Epidemiology,* 159, 882 – 890.

Tian, L., Cai, T. & Wei, L. J. (2009), "Identifying subjects who beneÞt from additional information for better prediction of the outcome variables," *Biometrics,* 65, 894 – 902.

Uno, H., Cai, T., Tian, L. & Wei, L. J. (2007), "Evaluating prediction rules for t-year survivors with censored regression models," *Journal of the American Statistical Association,* 102, 527 – 537.

Uno, H., Tian, L., Cai, T, Kohane, I. S. & Wei, L. J. (2009), "Comparing Risk Scoring Systems Beyond the ROC Paradigm in Survival Analysis," Harvard University Biostatistics Working Paper Series. Working Paper 107. http://www.bepress.com/harvardbiostat/paper107.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., et al. (2002), "Gene expression profiling predicts clinical outcome of breast cancer," *Nature,* 415, 530–6.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W.,

Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H. & Bernards, R. (2002), "A Gene-Expression Signature as a Predictor of Survival in Breast Cancer," *The New England Journal of Medicine,* 347, 1999 – 2009.

van der Vaart, A. W. & Wellner, J. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics,* Springer-Verlag, New York.

Wand, M., Marron, J. & Ruppert, D. (1991), "Transformation in density estimation (with comments)," *Journal of the American Statistical Association,* 36, 343 – 361.

Ware, J. H. (2006), "The limitations of risk factors as prognostic tools," *The New England Journal of Medicine,* 355, 2615 – 2617.

Yu, K. & Jones, M. C. (1998), "Local linear quantile regression," *Journal of the American Statistical Association,* 93, 228 – 237.

Zhou, X.H. and Obuchowski, N.A. & McClish, D.K. (2002), "Statistical methods in diagnostic medicine," *Wiley Interscience, New York.*

Table 1. *Estimates of regression parameters for Cox's models with breast cancer data*

	without gene score	with gene score
	Est (SE)	Est (SE)
Age/10 [yrs]	-0.47 (0.17)	-0.57 (0.18)
Diameter of tumor [cm]	0.19 (0.11)	0.18 (0.12)
Lymph nodes	0.00 (0.08)	-0.01 (0.08)
Grade = 2 vs 1	1.00 (0.35)	0.74 (0.35)
Grade = 3 vs 1	1.11 (0.35)	0.66 (0.37)
Vascular invasion 1-3 vs 0	0.08 (0.37)	-0.10 (0.37)
Vascular invasion > 3 vs 0	0.81 (0.62)	0.64 (0.63)
Estrogen Status=Positive	-0.39 (0.23)	-0.16 (0.24)
Chemo or Hormonal =Yes	-0.54 (0.33)	-0.49 (0.33)
Mastectomy=Yes	0.13 (0.21)	0.21 (0.22)
Gene score		2.43 (0.67)

Figure 1. Scatterplots of the risk scores with and without bilirubin for subjects whose survival times are not censored by Year 5

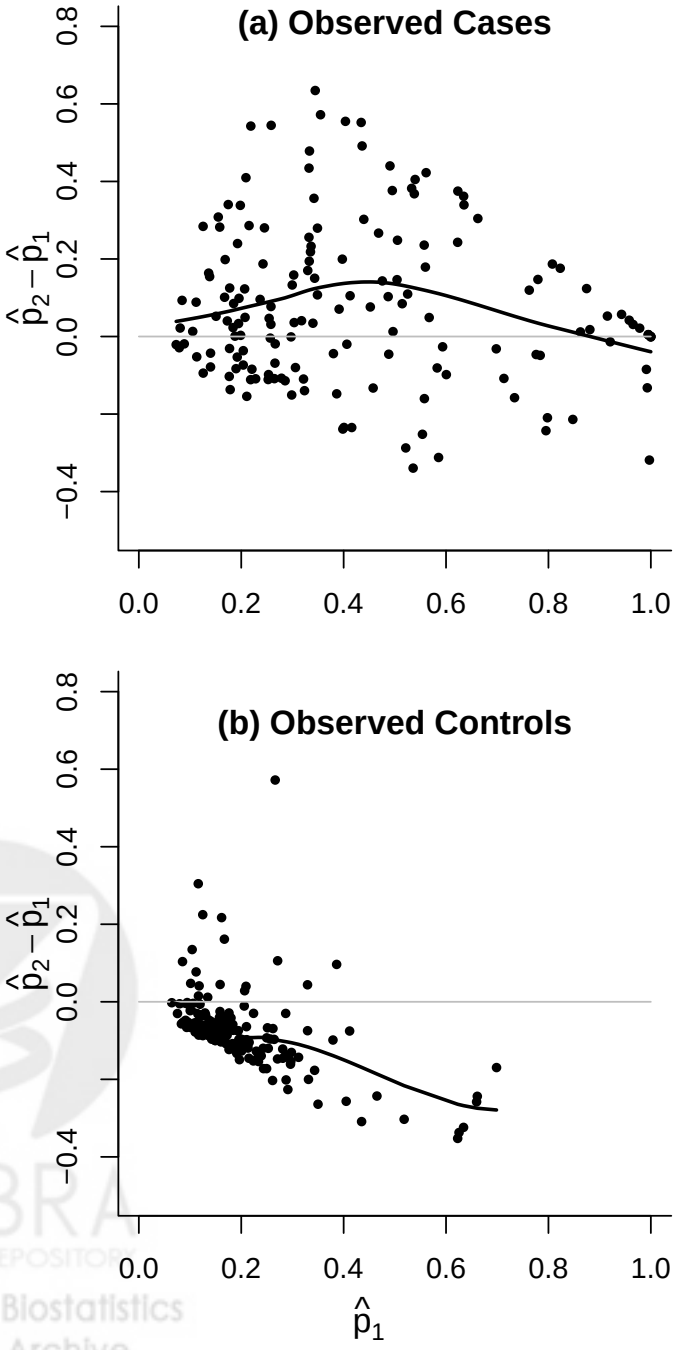

Figure 2. Estimates of cumulative mortality rates with Mayo PBC survival time data

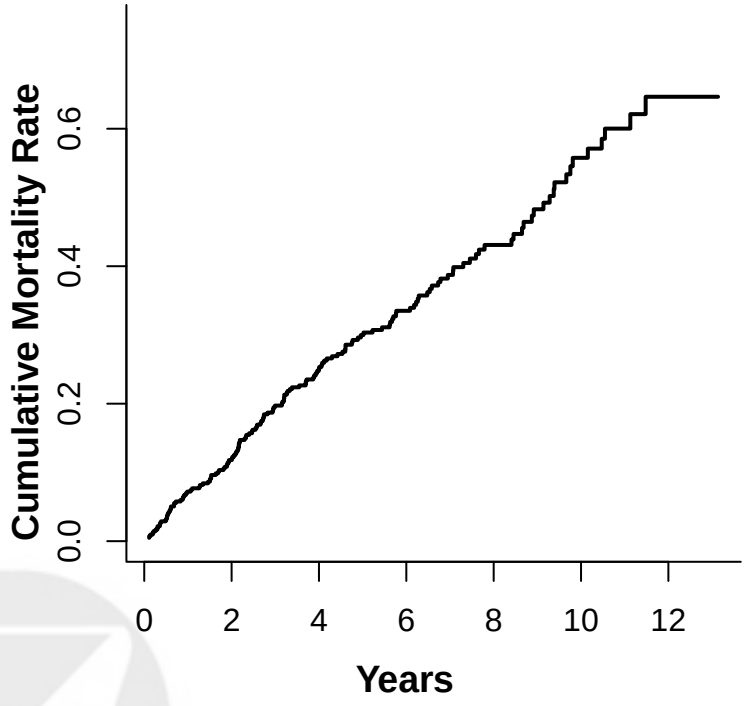

Figure 3. Estimated quartiles for the conditional distributions for the differences between the risk scores with and without bilirubin at Year 5. (a) For subjects who would die by Year 5; (b) For subjects who would survive beyond Year 5; (c) Estimated density function of the score without using bilirubin.

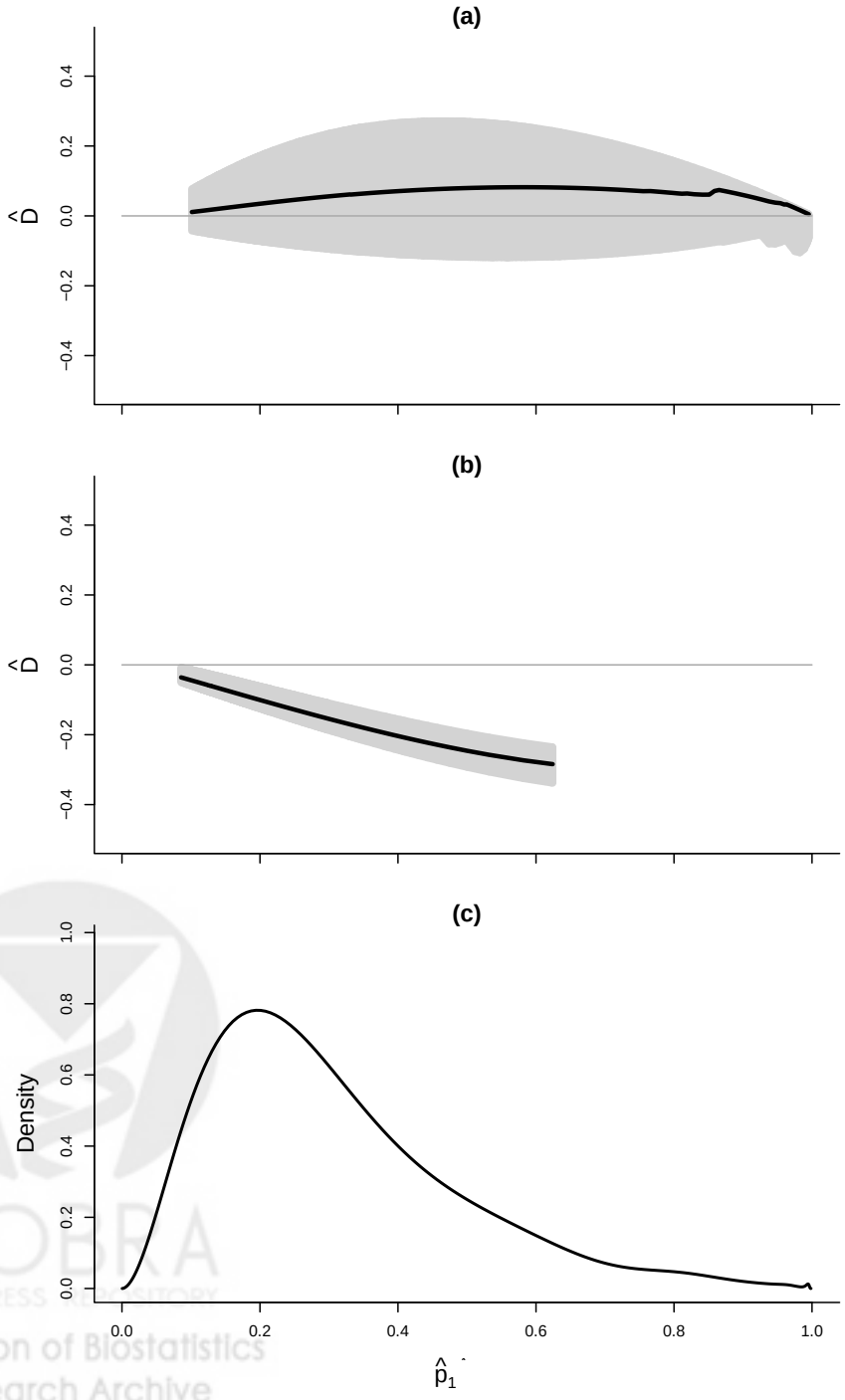

Figure 4. Estimated median curves for the conditional distributions for the differences between the risk scores with and without bilirubin for various sets of time intervals I_0 and I_1. (a) $I_1 = (0, 4], I_0 = (4, \infty)$; (b) $I_1 = (0, 4], I_0 = (5, \infty)$; (c) $I_1 = (0, 5], I_0 = (6, \infty)$; and (d) $I_1 = (0, 5], I_0 = (8, \infty)$.

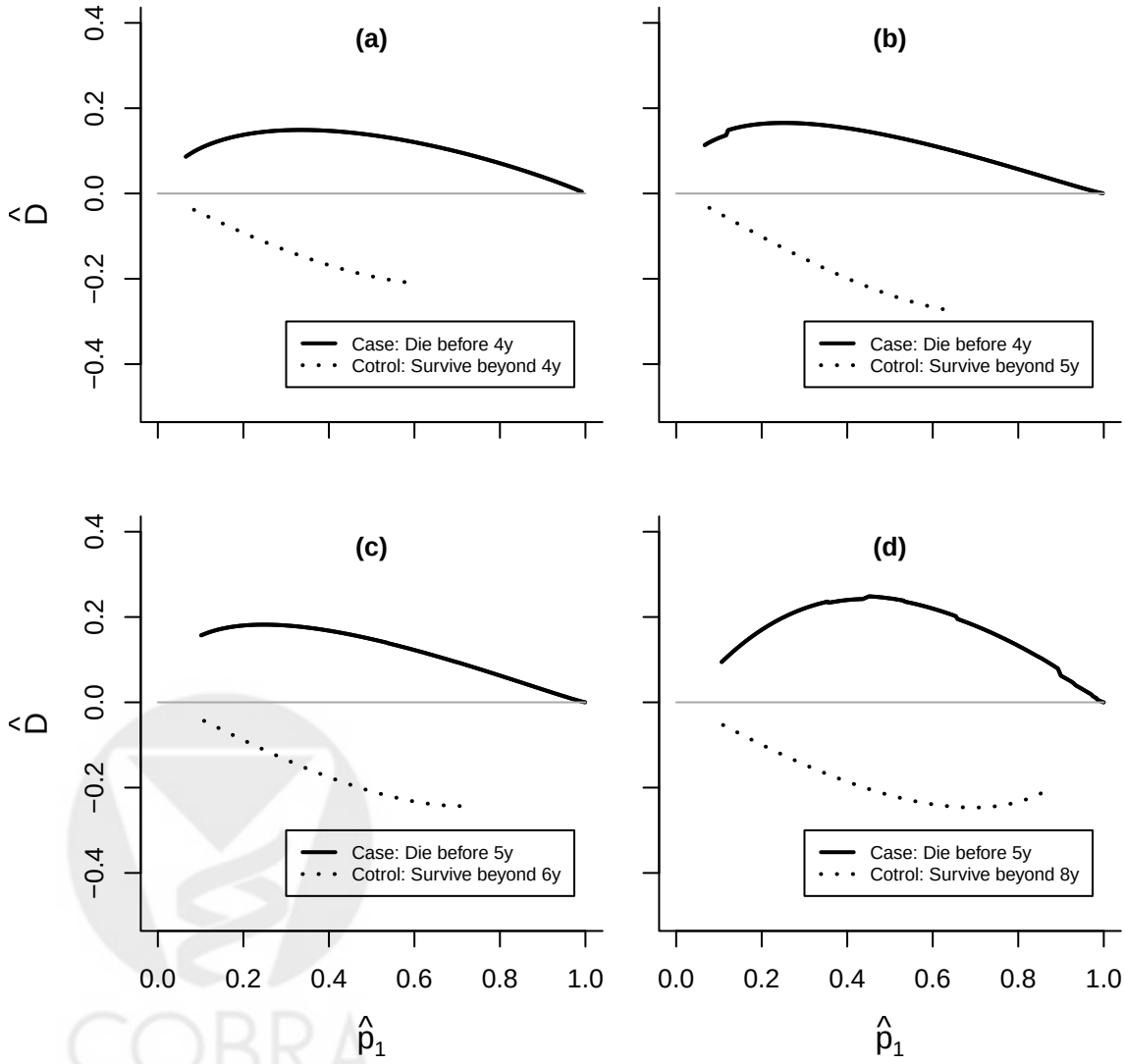

Figure 5. Estimates of cumulative mortality rates with breast cancer survival data

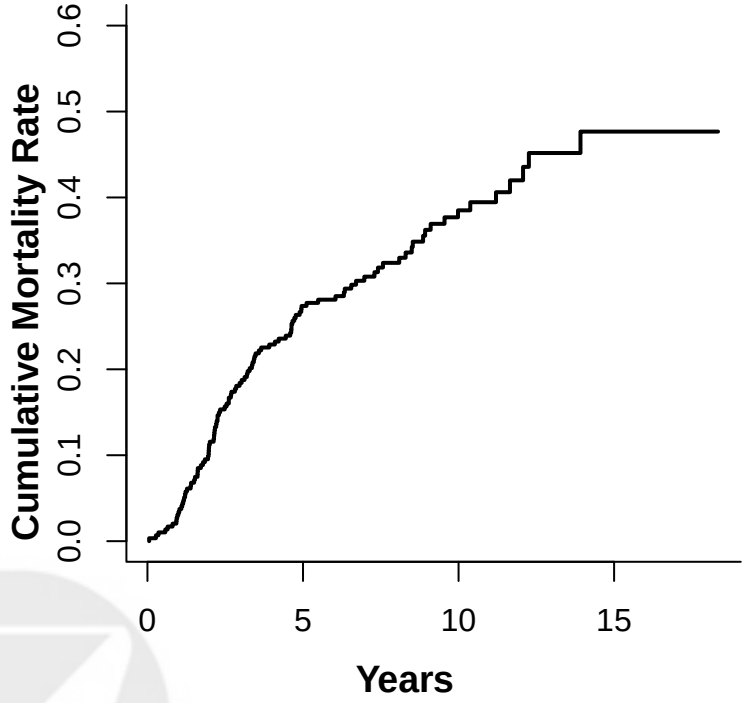

Figure 6. Estimated quartiles for the conditional distributions for the differences between the risk scores with and without gene score at Year 10. (a) For subjects who would die by Year 10; (b) For subjects who would survive beyond Year 10; (c) Estimated density function of the score without using gene score.

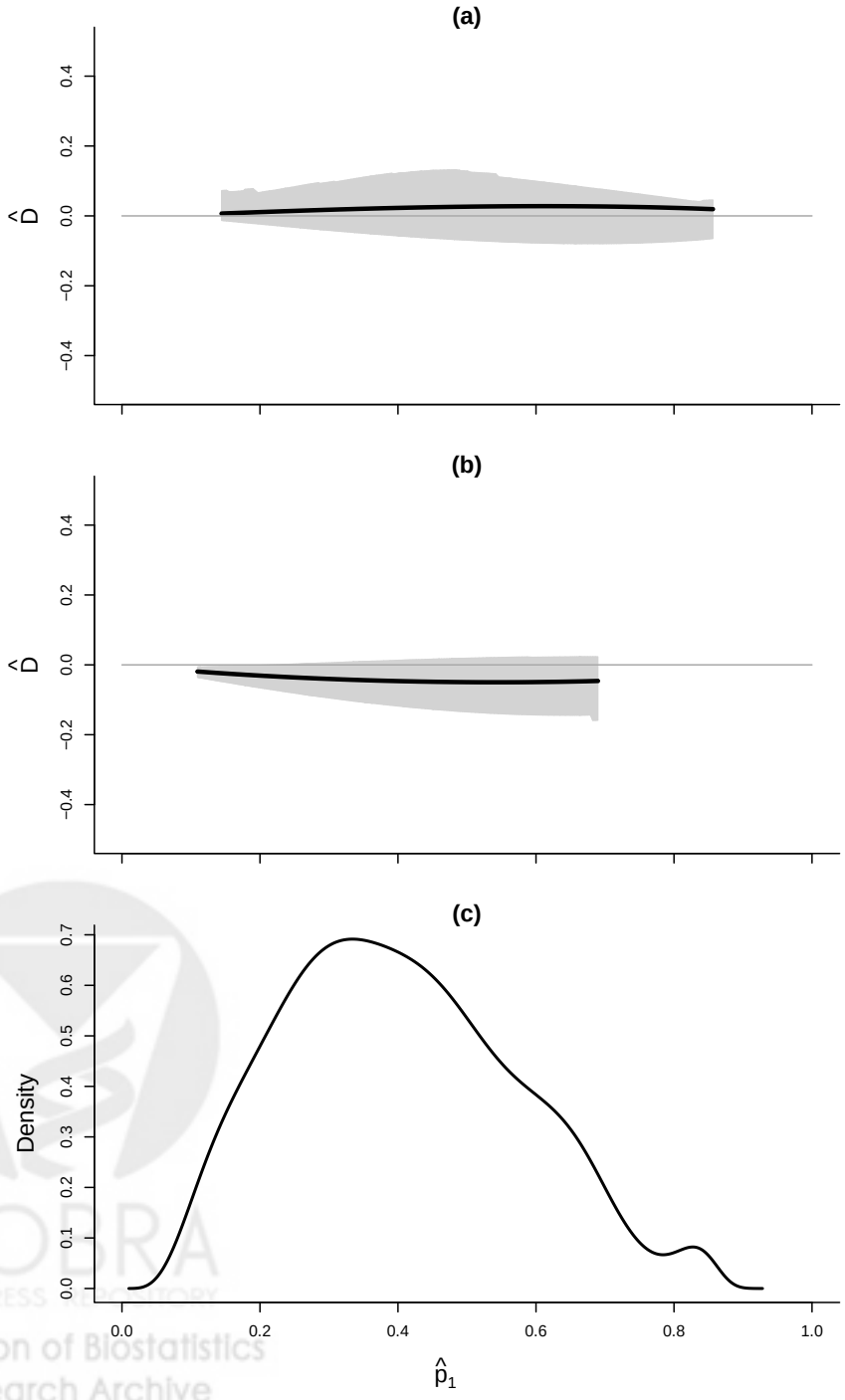

Figure 7. Estimated median curves for the conditional distributions for the differences between the risk scores with and without gene score for various sets of time intervals I_0 and I_1. (a) $I_1 = (0, 3], I_0 = (3, \infty)$; (b) $I_1 = (0, 3], I_0 = (8, \infty)$; (c) $I_1 = (0, 3], I_0 = (4, \infty)$; and (d) $I_1 = (0, 7], I_0 = (8, \infty)$.